Dual Linear Regression Based Classification for Face Cluster Recognition

Liang Chen
University of Northern British Columbia
Prince George, BC, Canada V2N 4Z9
chen.liang.97@gmail.com

Abstract

We are dealing with the face cluster recognition problem where there are multiple images per subject in both gallery and probe sets. It is never guaranteed to have a clear spatio-temporal relation among the multiple images of each subject. Considering that the image vectors of each subject, either in gallery or in probe, span a subspace; an algorithm, Dual Linear Regression Classification (DLRC), for the face cluster recognition problem is developed where the distance between two subspaces is defined as the similarity value between a gallery subject and a probe subject. DLRC attempts to find a "virtual" face image located in the intersection of the subspaces spanning from both clusters of face images. The "distance" between the "virtual" face images reconstructed from both subspaces is then taken as the distance between these two subspaces. We further prove that such distance can be formulated under a single linear regression model where we indeed can find the "distance" without reconstructing the "virtual" face images. Extensive experimental evaluations demonstrated the effectiveness of DLRC algorithm compared to other algorithms.

1. Introduction

> [Rules of Reasoning in Philosophy] Rule I:
> "We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances. "
> –Sir Isaac Newton

This paper deals with the following face cluster recognition problem: Given a galley set consists of a number of face image clusters, each containing the images of a known subject/identity; a probe cluster to be recognize contains a number of images of one subject; We are to match the probe cluster with the gallery set in order to determine the identity of the probe subject. This should be taken to be a typical image set based face recognition problem. However, in literature, it seems that the image based face recognition has been fallen under the category of video based face recognition. For example, most of the papers under the titles of image set based face recognitions, such as [10, 11, 28, 9, 27, 30, 5, 31, 1], use only benchmarks of video databases for the evaluation of their approaches. Therefore, we prefer call the problem we focus on "face cluster recognition" problem.

Video faces usually carry temporal relationships among image frames, and we usually are able to extract a large amount of images from a video clip. In face cluster recognition, it is never guaranteed to have a clear spatio-temporal relation among the images in one face cluster. For example, different face images in a cluster for one subject may be taken under different illumination conditions, with different poses and with different resolutions. We also require that the numbers of face images be much smaller than the number of pixels in an image, when two face clusters are to be matched. It is easy to know that there are many potential applications for the face cluster recognition problem. An example is for the integration of the documents of suspects in multiple law enforcement departments, where each suspect leaves a few images in the documents of one law enforcement department but a suspect is usually under different names/identifications in different departments; the first task to merge the documents is matching the face clusters. For real time recognition tasks such as airport surveillance systems, an ideal system should be able to perform face recognition without waiting to get lengthy videos with enough detectable face frames; in such situations, we can always expect that the extracted face frames be much smaller than the number of pixels in an image.

While there are a few approaches under the titles of image set recognitions focus on the estimations of parameters for representing image sets with certain distributions [14, 21], most of the work related fall under the category of non-parametric approaches where image sets are usually represented as a linear or nonlinear subspaces [28, 10]. We are intended to develop an non-parametric approach based on the idea of Linear Regression Classification (LRC) for still face probe recognitions on a gallery with multiple im-

ages per subject.

The simple but efficient linear regression based classification (LRC) approach was developed by Maseem, Togneri and Bennamoun [17]. LRC approach independently represents a downscaled probe image by the linear combination of the downscaled images of each subject in the gallery; the residual error of the representation is used to estimate the similarity between the probe and the cluster of faces of the subject. It is easy to understand that LRC approach belongs to the category of heuristic approaches, which includes SP (Sparse Representation ([29])), PCA (Principal Component Analysis [24]), S-LPP (Spatially Smooth Subspace Learning based Locality Preserving Projection ([4])), SRDA (Spectral Regression Discriminant Analysis ([2]), OLPP (Orthogonal Locality Preserving Projections ([3])) and Naseem et al. [17] has demonstrated its efficacy by extensive comparative studies with a few known state-of-the-art approaches of its same category.

In this paper, we propose a dual linear regression based classification (DLRC) algorithm to generalize the LRC approach for the face cluster recognition problem. When comparing two clusters of face images, we define the similarity between two clusters as the shortest distance between the subspaces each spanned from the face images of one cluster. In order to do so, DLRC attempts to find a "virtual" face image located in the intersection of the subspaces spanning from both clusters of downscaled face images (See Figure 1). The "distance" between the "virtual" face images reconstructed from both subspaces is then taken as the distance between these two subspaces. We further prove that such distance can be formulated under a single linear regression model where we indeed can find the "distance" without reconstructing the "virtual" face images.

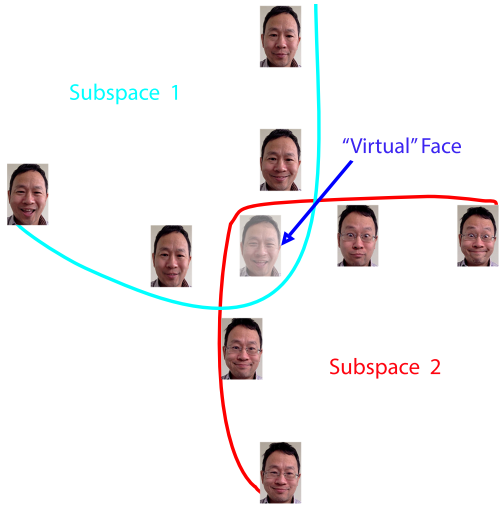


Figure 1. A "virtual" face in the interection of two subspaces

The rest of the this paper is organized as follows. We propose the DLRC concept and algorithm in Section 2. This is followed by extensive experiments in Section 3. The conclusion and discussion are given in Section 4.

Parts of the program codes and datasets are available via: http://web.unbc.ca/~chenl/DataCode.html.

2. Dual Linear Regression based Classification for Face Cluster Recognition

2.1. Math Derivations

Let two clusters of (downscaled) face images be represented by

$$X = [x_1\ x_2\ \cdots\ x_m], \tag{1}$$

and

$$Y = [y_1\ y_2\ \cdots\ y_n]. \tag{2}$$

where x_i, $i = 1, 2, \cdots, m$, and y_j, $j = 1, 2, \cdots, m$ are column vectors of size $1 \times ab$, each representing the an downscaled image of size $a \times b$. We also require that $ab \geq m+n$ in order to ensure that the number of pixels of an image is no less than the total number of images in these two clusters.

An image located in the subspace spanned by the column vectors of either X or Y should be a linear combination of these column vectors. The task for locating a "virtual" face the intersection of both subspaces is to find V and $\alpha = (\alpha_1\ \alpha_2\ \cdots\ \alpha_m)^T$, $\beta = (\beta_1\ \beta_2\ \cdots\ \beta_n)^T$ such that $V = X\alpha$ and $V = Y\beta$.

However, it is easy to know that there is a trivial solution for the task where $V = \mathbf{0}$, $\alpha = \mathbf{0}$ and $\beta = \mathbf{0}$. Obviously this is not what we want.

Considering that we can have all downscaled images standardized into unit vectors, we further require that

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 1. \tag{3}$$

Thereafter, we are to find $\hat{\alpha} = (\alpha_1\ \alpha_2\ \cdots\ \alpha_{m-1})^T$, $\hat{\beta} = (\beta_1\ \beta_2\ \cdots\ \beta_{n-1})^T$, such that

$$\begin{aligned} V &= [\hat{x}_1\ \hat{x}_2\ \cdots\ \hat{x}_{m-1}]\hat{\alpha}^T + x_m \\ &= [\hat{y}_1\ \hat{y}_2\ \cdots\ \hat{y}_{n-1}]\hat{\beta}^T + y_n \end{aligned} \tag{4}$$

where $\hat{x_i} = x_i - x_m$, $i = 1, 2, \cdots, m-1$ and $\hat{y_j} = y_j - y_n$, $j = 1, 2, \cdots, n-1$. Assuming that there is an approximate solution $\gamma' = (\gamma'_1\ \gamma'_2\ \cdots\ \gamma'_{m+n-2})^T$ for the Equation

$$y_n - x_m = \widehat{XY}\gamma, \tag{5}$$

where

$$\widehat{XY} = [\hat{x_1}\ \hat{x_2}\ \cdots\ \hat{x}_{m-1}\ -\hat{y_1}\ -\hat{y_2}\ \cdots -\hat{y}_{n-1}].$$

The reconstructed "virtual" faces from subspaces of X and Y are:

$$V'_X = [\hat{x_1}\ \hat{x_2}\ \cdots\ \hat{x}_{m-1}](\gamma'_1\ \gamma'_2\ \cdots\ \gamma'_{m-1})^T + x_m, \tag{6}$$

and

$$V'_Y = [\hat{y_1}\ \hat{y_2}\ \cdots\ \hat{y}_{n-1}](\gamma'_m\ \gamma'_{m+1}\ \cdots\ \gamma'_{m+n-2})^T + y_n. \tag{7}$$

Therefore, the difference between V'_X and V'_Y is

$$V'_Y - V'_X = (y_n - x_m) - \widehat{XY}\gamma'. \tag{8}$$

Equation 8 shows an important conclusion:

Conclusion: The difference between the reconstructed the "virtual" images from both subspaces is equivalent to the residual error of the solution of Equation 5.

Therefore, we can use the residual errors of the regression solution for Equation 5 to estimate the similarity between the two subspaces, the smaller the better.

Depending on the qualities of face images, γ can be estimated using either least squares (LS) or least trimmed squares (LTS) objective functions. We can expect that, if the images are occluded, we should use least trimmed squares (LTS) as the objective function, otherwise, least squares (LS) should be the objective function.

Assuming $r(\gamma) = (y_n - x_m) - \widehat{XY}\gamma$ and $r(\gamma) = (r_1(\gamma)\ r_2(\gamma)\ \cdots\ r_{ab}(\gamma))'$, let $\{|r(j)(\gamma)|, 1 \le j \le ab\}$ denote the set of increasingly ordered absolute values of the residuals $\{|r_1(\gamma)|, |r_2(\gamma)|, \cdots, |r_{ab}(\gamma)|\}$, the LS and LTS based estimations are to find the solution for the Equation:

$$\arg\min_{\gamma} \sum_{j=1}^{h} r(j)(\gamma), \tag{9}$$

where $h = ab$ for LS and $h < ab$ for LTS.

When $\widehat{XY}^T\widehat{XY}$ is not singular, for LS estimation, the best solution of γ can be given by [20]:

$$(\widehat{XY}^T\widehat{XY})^{-1}\widehat{XY}^T(y_n - x_m). \tag{10}$$

For LTS estimation, the fast-LTS algorithm by Rousseeuw and Van Driessen [19] can be used to find approximate and generally sufficiently accurate solutions.

2.2. Algorithm

The entire DLRC algorithm for finding the identity of a probe cluster against a galley set consisting of subject clusters, each of which contains the images of one subject, works as follows:

1. Downsample and Normalize all gallery and all probe images into size $a \times b$.
2. Construct a matrix Y as shown in Equation 2 using the probe images.
3. For each subject in the gallery, construct a matrix X as shown in Equation 1.
   - 3.1. Find the estimation of γ for the Equation 5, with respect to the objective function as shown in Equation 9.
   - 3.2. Use the sum of residuals as shown in 9 as the similarity between the subject and the probe.
4. Compare the similarities between all subjects and the probe, choose the subject closest to the probe as the answer.

Limitation: In order to find a solution, we require that $\widehat{XY}^T\widehat{XY}$ in Equation 10 be not singular. Therefore, we should require that the total number of images in two clusters should be smaller than the pixel number in the downscaled images when we computing the distance between these two clusters of images. In practice, to avoid a singularity in computing with a real computing machinery which has fixed working precision (typically double precision), each images in each cluster should be significantly smaller than half of the number of pixels in a downscaled image.[1]

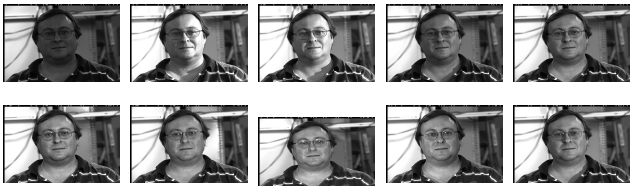
3. Experiments

We have carried out the face cluster recognition experiments on face clusters from CMU PIE and LWF databases. Although as we have mentioned in Section 1 that face cluster recognition is significantly different from video based face recognition, we also carry a few experiments on Hongda/HCSD, CMU Mobo and YouTube Celebrities datasets to demonstrate its applicabilities in video based face recognition.

3.1. Experiments on PIE database

The following two sets of PIE face images [22] are used:
Gallery Set: contains the images taken by camera ID C27, with flash IDs 00 and 02 to 10 (flash ID 00 corresponds to no flash), the background (neutral illumination) being turned on and the subjects wearing glasses if and only if they normally do. As a total, there are 680 images of 68 subjects in this gallery set, each subject having 10 images.
Probe Set: Contains the images of each individual taken by camera IDs C02, C09, C11, C14, C22, C25, C29, C31, C34, and C37, under neutral expressions and the subjects wearing glasses if and only if they normally do. There are 680 images of 68 subjects in the Probe Set, each subject having 10 images. Note that, camera IDs C05, C07 and C27 take upfront or close to upfront view face photos – these are not included in this probe set. The original images (without manually rotated nor shifted) of size 486×640 pixels are used in our experiments. Figure 2 represents the gallery and probe images of a subject in our experiments.

[1] It may be possible to avoid a singularity by adding a regularization term which has to be determined by further investigations.



(1) Gallery Images of A Subject

(2) Probe Images of A Subject

Figure 2. Sample Images for PIE Experiments

Image Sizes	DLRC	LRC+NN	SR + NN
10×10	16.18%	13.24%	8.82%
15×10	20.59%	13.24%	10.29%
30×30	19.12%	8.82%	11.76%

MDA	AHISD	CHISD	SANP
4.41%	16.18%	20.59%	23.53%
5.88%	22.06%	16.18%	20.59%
–	16.18%	14.71%	17.63%

Table 1. Accuracies of PIE Experiments

We are to match each probe cluster with the gallery set of face clusters. Note again, each cluster contains the images of one subject. We have done the experiments using our DLRC approach on downscaled images of sizes 10×10, 15×10 and 30×30. We use the least squares (LS) objective function. We have also carried out the experiments with the same sized images using the LRC approach + Nearest Neighbor (NN) proposed by [17], where LRC approach finds the distance between a probe image and a gallery cluster, NN strategy is used to the image in a probe cluster which is closest to a gallery cluster, and the distance between this image and the gallery cluster is taken to be the distance between the probe cluster and the gallery cluster. The MDA (Manifold Discriminant Analysis) [26], SANP (Sparse approximated nearest points) [11], AHISD (Affine Hull based Image Set Distance) and CHISD (Convex Hull based image Set Distance) ([5]), and SR (Sparse Representation ([29])) + NN, are also applied in this experiments. All the results are summarized in Table 1, where "-" represents a case that a singularity was reached so that the computing was terminated by the machine.

3.2. Experiments on LFW Set

To further illustrate the performances of our DLRC approach, we have carried out the experiments in "Labeled Faces in the Wild" (LFW). "Labeled Faces in the Wild" (LFW) [12] is available via the LFW official site http://vis-WWW.cs.umass.edu/lfw/results.html. All the face images in LFW were taken in unconstrained environments, exhibiting " 'natural' variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality" [12]. We use LFW-a version of the images (the images aligned with a commercial face alignment software) [23]. The LFW-a version images are of size 250×250, we first crop them into images of size 90×78 (by removing 88 pixel margins from top, 72 from bottom, and 86 pixel margins from both left and right sides). Note that, there were many errors in the alignment; we just keep them as they were (so some of the final cropped faces indeed are not correctly aligned). We select all the subjects in LFW who have at least 20 pictures. As a total, there are 62 such subjects. We use the first 10 images of each subject as the training images and the last 10 as the probe images.

Figure 3 represents the gallery images and the probe images of a subject in our experiments.

(1) Samples of Gallary Images

(2) Samples of Probe Images

Figure 3. Sample Images for LFW Experiments

We have done the experiments using our DLRC approach on downscaled images of sizes 10×10, 15×10 and 30×15. We use the least squares (LS) as the objective function. We have also carried out the experiments with the same sized images using the LRC approaches proposed by [17] +NN (nearest neighbor) strategy, and the SR (Sparse Representation ([29])+NN, MDA (Manifold Discriminant Analysis) [26], SANP (Sparse approximated nearest points) [11], AHISD (Affine Hull based Image Set Distance) and CHISD (Convex Hull based image Set Distance) ([5]), the results are summarized in Table 2.

Image Sizes	DLRC	LRC+NN	SR + NN
10×10	8.06%	22.58%	14.52%
15×10	4.84%	16.13%	11.29%
30×15	1.61%	12.90%	3.23%

MDA	ASIHD	CSIHD	SANP
6.45%	12.90 %	9.68%	14.52%
4.84%	4.84%	6.45%	7.45%
19.35%	6.45%	3.23%	4.84%

Table 2. Error Rates of LWF Experiments

It is clear again that the DLRC approach performs much better than all other algorithms of same categories.

3.3. Experiments on Honda/UCSD database

The Honda/UCSD data [15] contains 59 video clips of 20 subjects, all but one have at least 2 videos. 20 videos are called training videos and the other 39 test videos. The lengths of videos vary from 291 to 1168 frames. Following the settings of [5, 11, 10, 26, 31], we use Viola-Jone cascaded face detector [25] (implemented in Matlab, which was planted from Open CV standard implementation, with all default settings) to extract faces frame by frame in each video. Histogram equalization is employed to reduce the illumination variations. Then resize them into 10×10 and 20×20 gray images. Example of extracted faces are shown in Figure 4, each row represents a set of faces extracted from one video file. Note that, there were a few false detected "faces" (some of them are actually not faces), we just keep them in the extracted face clusters as they were.

This dataset has been used extensively for image-based face recognition, the accuracy has reached 100% or close to 100%. Therefore, researchers have turned to experiment on the settings using a small amount frames. The results of a few the-state-of-the-art algorithms, including DCC [14], MMD[28], MDA[26], SANP and KSANP[11], AHISD and CHISD[5], RNP[31] on this dataset using the first 50 frames are recently reported in [11] and [31]. We now carry the experiments using first 40 frames for image sizes 10×10, and 50 frames for image size 20×20. (Note: As we mentioned in Section 2.2 , we require the number of images in each face set is (significantly) less than half of number of pixels in a image; this is the reason we cannot choose "50" for image size 10×10.) The results are summarized in Table 3.[2]

It is easy to see that, our DLRC is able to get better accuracies with smaller image sizes and less frames.[3] We should

Figure 4. "Face" samples extracted from Hongda/UCSD Dataset

Methods	First 50 Frames
	Image Size 20×20
DCC	70.92%
MMD	69.32%
MDA	82.05%
AHISD	87.18%
CHISD	82.05%
MSM	74.36%
SANP	84.62%
KSANP	87.18%
RNP	87.18%

DLRC	
First 40 Frames	First 50 Frames
Image Size 10×10	Image Size 20×20
89.74%	92.31%

Table 3. Accuracies on the Hongda/UCSD Dataset

note here that, we can indeed apply DLRC to the situations where the images in a cluster is much larger than the image pixel number: randomly select a sub-cluster of images from each cluster, run the program on the small sized sub-clusters; run the code a few times, each on different sub-clusters, and vote on the results. We are able to get 100% for DLRC with this strategy, however, we feel it is unfair to

[2]The accuracies of DCC, MMD, MDA, AHISD, CHISD, SANP and KSANP in Table 3 were copied from Table 2 in [11], the accuracies of MSM and RNP were reported in [31].

[3]We notice that it seems that the face sets extracted from Honda/UCSD by different research groups are not always equivalent, even though the extraction approaches used are all claimed to be the Viola-Jone detector. For example, although [28] claimed that they also extracted the faces from Honda/UCSD via Viola-Jone detector (therefore their extracted face sets and our version carry roughly the same amount of information since both were extracted from the same video sets), a careful exam can find that the face sets extracted by [28] (available via their authors' website) are not exactly the same as the sets we fully automatically extracted (eg. the version of [28] does not contain any false detected faces such as the second image in first row, nor the fifth and the ninth in the fifth row of Figure 4). Therefore, the comparisons of the accuracies in Table 3 should not be taken to be very precise.

compare such results with others, since this does not agree with the common protocol.

3.4. Experiments on CMU Mobo database

The CMU Mobo dataset [8] contains the video sequences of 25 subjects walking on a treadmill. All but the last one have 4 different videos collected in four walking patterns, namely, holding a ball, fast walking, slow walking and incline walking. Usually the videos of the first 24 subjects are used in the experiments of image set based face recognitions. For this video set, the standard protocol uses Viola-Jones algorithm to extract faces from videos, the-state-of-the-art works usually employ LBP approach to extract unified LBP histogram features (with circular (8,1) neighborhood, 8×8 blocks) from the face images before the recognition takes place (eg. [5, 31, 11]). We here use directly the processed LBP histogram features of face images available via the website of the authors of [5].

Recent researches on this dataset have reached 98+% accuracies, and therefore, researchers are competing their algorithms on a small amount of frames [31]. To be consist with Section 3.3, our experiments use only 50 frames. The results reported at [28] are the averages on "10 random splits", where in each split, randomly one video from one subject is chosen for training and the rest for test. In order to have a fair comparison, we generate the averaged accuracy with 1000 runs, each with "10 random splits". The averaged accuracy of each 10-random-splits-experiment is also computed, and the worst and best cases of such 10-random-split averaged accuracies are also recored. The results, as well as the tested results of a few the-state-of-the-art algorithms, including DCC [14], MMD[28], MDA[26], SANP and KSANP[11], AHISD and CHISD[5], RNP[31], available at [31], are summarized in Table 4. [4]

The average accuracies and standard deviations of the worst and the best cases of 10-random-split experiments, as shown in Table 4, clearly indicate that our algorithm is comparable to any of the-state-of-the-art algorithms for this experiment.

3.5. Experiments on YouTube Celebrities database

The YouTube data [13] were collected from YouTube, it contains 1,910 video sequences of 47 celebrities. The video length varies from 7 to 350 frames, most of which are low quality low resolution. It has been observed [11] that Viola-Jone cascaded face detector [25] often fails on this dataset; and it is easy to know an initial assignment for many of the videos, where there are two or more individuals showing up, is absolutely necessary in order to automatically crop the faces of the correct individual from the them. Therefore we follow [11]: use the IVT (Incremental Learning for Visual Tracking) [18] tool to track and extract faces frame by frame using the information of the cropped face in the first frame available via the dataset website [5]. We use all the default settings of the IVT tool. Figure 5 shows a few examples of the tracked and cropped faces from this dataset. We resized the faces into two versions, 20×20 and 10×10.

Methods	10 random splits
DCC	$82.1\% \pm 2.7\%$
MMD	$90.1\% \pm 2.3\%$
MDA	$86.2\% \pm 2.9\%$
AHISD	$91.6\% \pm 2.8\%$
CHISD	$91.2\% \pm 3.1\%$
MSM	$84.3\% \pm 2.6\%$
SANP	$91.8\% \pm 3.1\%$
RNP	$91.9\% \pm 2.5\%$

DLRC
Ave. $\pm$ std. of all 10,000 Splits $91.60\% \pm 2.78\%$
Ave. $\pm$ std. of the means of 1,000 runs of "10-Splits" experiments $91.60\% \pm 0.91\%$
Best "10-Splits": $94.17\% \pm 1.50\%$
Worst "10-Splits": $88.61\% \pm 3.50\%$

Table 4. Average Accuracies & Standard Deviations on the CMU Mobo Dataset

Figure 5. "Face" samples extracted from YouTube Celebrities

We follow the five-fold cross validation setting of the experiences of [10, 11], where Hu et al. has carried experiments on the IVT tracked faces of the YouTube Celebrities: Partition the videos of each subjects equally into the 5 folds,

[4] The accuracies of DCC, MMD, MDA, AHISD, CHISD, MSM and SANP in Table 4 were copied from Table 3 in [31].

[5] http://seqam.rutgers.edu/site/index.php?option=com_content&view=article&id=64&Itemid=80

each contains 9 videos per subject[6]. In each fold, 3 videos are randomly selected for training, the rests for testing.

Our DLRC uses only 40 frames of each video clip (when an clip has than 40 images, use all) for experiments on this database[7]. The results reported in [11] are the averages and standard deviations of one run of such a five-fold-cross-validation. We report our results on 1000 runs of "five-fold-cross-validations": Each run gets a mean and a standard deviation, we report the best and the worst cases of the 1000 runs; we also report the average and the standard deviation of the 1000 means (each represents one run of "five-fold-validation"). Our results and the results reported in [10, 11] on the face image sets extracted from same video sets using the same IVT tool are shown in Table 5.[8] We can see that while the reported results of other approaches use all the frames of all subjects, our first 40 frame experiments on both image sizes have already reached the-state-of-the-art accuracy using all frames (10×10 version is slightly better than the best published result).

We use only first 40 frames and we reach the-state-of-the-art performances where all other approaches use all frames, we can expect that our approach is more valuable than others for applications in emergency real time tasks – since we can perform the recognition task without waiting for longer video chips.

3.6. Running Time Efficiency

To have a fair comparison for time efficiency tests, we run the experiments of DCC [14], MMD[28], MDA[26], SANP, AHISD and CHISD[5], and our DLRC, using only 50 frames on CMU Mobo dataset using the setting in Section 3.4. We use a Laptop with Windows 7, CPU i7 M620 2.67GHz. The results are shown in Table 6 (We only try to find the time complexity here, we do not adjust parameters to find best accuracy for each approach), where "/" indicates no training is involved. We can see that DLRC is the fastest – this is the also reason that we can get 1000 runs of random 10-splits/5-folds experiments: we first get pairwise similarities between all pairs of video clips, then we do required random partitions to get the results of all the 1000 runs.

4. Discussions and Future Work

We have developed a Dual Linear Regression based Classification (DLRC) algorithm for face cluster recognition. Experiments have demonstrated the efficiency of our DLRC algorithm compared to a few image set based face recognition algorithms: Section 3.6 shows that our DLRC is much faster than other algorithms when dealing with videos with same lengths. The experiment in Section 3.5 demonstrated that we can reach the-state-of-the-art results on real video images with less frames.

Considering the computing efficiency of DLRC in comparing to other approaches, DLRC could be also used for real time face recognitions, where we don't have to wait for a full and lengthy video frame before a recognition process is performed.

A limitation of the DLRC approach is that the number of images in a cluster should be smaller than a half of the number of pixels of an image. For a possible application on video based face recognition, we can partition the videos into a number of clusters, apply DLRC on each cluster and make final decision through a voting process.

Chen [7, 6] has established a theory that, for any algorithm for subjective pattern recognition tasks, such as

Methods	All Frames
	Image Size
	40×40
DCC	$53.90\% \pm 4.68\%$
MMD	$54.04\% \pm 3.69\%$
MDA	$55.11\% \pm 4.55\%$
AHISD	$60.71\% \pm 5.24\%$
CHISD	$60.42\% \pm 5.95\%$
SANP	$65.03\% \pm 5.74\%$
KSANP	$65.46\% \pm 5.53\%$

DLRC	
First 40 Frames	
Image Size 10×10	Image Size 20×20
Ave. $\pm$ Std. of all 5,000 folds	
$66.18\% \pm 4.34\%$	$65.55\% \pm 5.16\%$
Ave. $\pm$ Std. of the means of 1,000 runs of "5-fold-validation"	
$66.18\% \pm 1.04\%$	$65.55\% \pm 1.24\%$
Best "5-fold-validation"	
$69.29\% \pm 3.24\%$	$69.15\% \pm 3.76\%$
Worst "5-fold-validation"	
$63.19\% \pm 4.16\%$	$61.70\% \pm 5.12\%$

Table 5. Accuracies and Standard Deviations on the YouTube Celebrities Dataset

	DCC	MMD	MDA	AHISD	CHISD	SANP	**DLRC**
Training	21.31	40.34	1998.32	/	/	/	**/**
Testing	1.21	1.43	11.02	1.52	3.24	10.71	**1.15**

Table 6. Time Costs (seconds) for Experiments on Mobo dataset

[6]It was not very clear that how the image sets are partitioned for different situations. We confirmed with the authors of [10, 11]: When an individual has less than 45 image sets, they make sure that there are as less overlapping as possible; at the time when a subject has more than 45 image sets, then only first 9 of each group are kept after the image sets are partitioned into 9 groups.

[7]There are a few cases that $\widehat{XY}^T \widehat{XY}$ in Equation 10 get to close to singular when we choose 50 frames.

[8]The accuracies of DCC, NMD, MDA, AHISD, CHISD, SANP and KSANP in above table were copied from Table 4 in [11].

face recognition, an "Electoral College" version can always work better. For the cases of matching a face image with a set of face images, Naseem, Togneri and Bennamoun [17] have shown by experiments that, when the images contain large amount of occlusions, such as scarves, the performances can be improved if the following strategy, which has been called "Electoral College" by Chen et al. [7, 6], is taken: partition the images into blocks, use LRC approach in the corresponding blocks of all images, and then make the final decision by voting. We trust that we can further improve the performance of our DLRC approach by using the same strategy.

Acknowledgment. This work is supported by NSERC Discovery Grant (No. 261403-2011 RGPIN).

References

[1] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, and T. Darrell. Face recognition with image sets using manifold density divergence. *CVPR 2005*, vol. 1, pp.581–588. 1

[2] D. Cai, X. He, and J. Han. SRDA: An efficient algorithm for large scale discriminant analysis. *IEEE TKDE*, 20(1):1–12, January 2008. 2

[3] D. Cai, X. He, J. Han, and H.-J. Zhang. Orthogonal laplacianfaces for face recognition. *IEEE Trans. Image Processing*, 15(11):3608–3614, Nov. 2006. 2

[4] D. Cai, X. He, Y. Hu, J. Han, and T. Huang. Learning a spatially smooth subspace for face recognition. *CVPR 2007*. 2

[5] H. Cevikalp and B. Triggs. Face recognition based on image sets. *CVPR*, pp.2567–2573, 2010. 1, 4, 5, 6, 7

[6] L. Chen. Robustness instead of accuracy should be the primary objective for subjective pattern recognition research: Stability analysis on multicandidate Electoral College versus direct popular vote, *Computational Intelligence*, doi:10.1111/j.1467-8640.2012.00439.x, 2012. 7, 8

[7] L. Chen and N. Tokuda. A general stability analysis on regional and national voting schemes against noise —- Why is an electoral college more stable than a direct popular election? *Artificial Intelligence*, 163:47–66, 2005. 7, 8

[8] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA, June 2001. 6

[9] A. Hadid and M. Pietikainen. From still image to video-based face recognition: an experimental analysis. *The Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp.813–818, 2004. 1

[10] Y. Hu, A. Mian, and R. Owens. Sparse approximated nearest points for image set classification. *CVPR 2011*, pp.121–128. 1, 5, 6, 7

[11] Y. Hu, A. S. Mian, and R. Owens. Face recognition using sparse approximated nearest points between image sets. *IEEE TPAMI*, 34(10):1992–2004, 2012. 1, 4, 5, 6, 7

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 4

[13] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. *CVPR*, pp.1–8, 2008. 6

[14] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE TPAMI*, 29(6):1005–1018, 2007. 1, 5, 6, 7

[15] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *CVPR 2003*, vol. 1, pp.I–313–I–320. 5

[16] A. Martinez and R. Benavente. The ar face database. CVC Tech. Report 24, School of Electrical & Computer Engineering, Purdue University, West Lafayette, Indiana, 1998.

[17] I. Naseem, R. Togneri, and M. Bennamoun. Linear regression for face recognition. *IEEE TPAMI*, 32(11):2106–2112, 2010. 2, 4, 8

[18] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008. 6

[19] P. Rousseeuw and K. Van Driessen. Computing lts regression for large data sets. *Data Mining and Knowledge Discovery*, (12):29–45, 2006. 3

[20] G. A. F. Seber. *Linear Regression Analysis*. John Wiley & Sons, Inc., New York, 2003. 3

[21] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. *ECCV 2002*, pp.851–865. 1

[22] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression database. *IEEE TPAMI*, 25(12):1615–1618, 2003. 3

[23] Y. Taigman, L. Wolf, and T. Hassner. Multiple one-shots for utilizing class label information. *BMVC 2009*. 4

[24] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. 2

[25] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57:137–154, 2004. 5, 6

[26] R. Wang and X. Chen. Manifold discriminant analysis. *CVPR 2009*, pp.429–436. 4, 5, 6, 7

[27] R. Wang, H. Guo, L. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. *CVPR 2012*, pp.2496–2503. 1

[28] R. Wang, S. Shan, X. Chen, and W. Gao. Manifold-manifold distance with application to face recognition based on image set. *CVPR 2008*, pp.1–8. 1, 5, 6, 7

[29] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI*, 31(2):210–227, 2009. 2, 4

[30] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao. Set based discriminative ranking for recognition. *ECCV 2012*, pp.497–510. 1

[31] M. Yang, P. Zhu, L. Van Gool, and L. Zhang. Face recognition based on regularized nearest points between image sets. *FG 2013*, pp.1–7. 1, 5, 6